International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN (P): 2249-6831; ISSN (E): 2249-7943
Vol. 8, Issue 3, Aug 2018, 1-8

LEUKEMIA PREDICTION USING RANDOM FOREST ALGORITHM

NORA NAIK[1], PETER BRAGANZA[2], AARON CORDEIRO[3], RONAN D'SOUZA[4]& RAYSTER FERNANDES[5]

[1]*Assistant Professor, Department of Computer Engineering, Agnel Institute of Technology & Design, Goa, India*

[2,3,4,5]*Students, Department of Computer Engineering, Agnel Institute of Technology & Design, Goa, India*

ABSTRACT

The analysis of white blood cells in microscopy images allows the evaluation of hematic pathologies such as Acute Lymphoblastic Leukemia (ALL). Classification of white blood cells (WBC) is usually done manually by experienced hematologists. The efficiency and accuracy of this process depend on the skill and experience of the operator as well as his state of mind. On account of these reasons, the outcome of the classification may be undesirable. In this paper, we present a methodology for fast automated segmentation of white blood cells from blood image sample. The focus lies in the classification algorithms viz. Random Forest and k Nearest Neighbor (kNN), which are used to classify cells as blast cells or not. The classification model is built from the features extracted from the blood smear images using the various image processing techniques.

KEYWORDS*: *Acute Lymphoblastic Leukemia (ALL), Image Processing, KNN Classifier & Random Forest Classifier

Received: Apr 20, 2018; **Accepted:** May 11, 2018; **Published**: Jun 30, 2018; **Paper Id**.: IJCSEITRAUG20181

Original Article

INTRODUCTION

Leukemia is cancer of the blood cells that mostly originate in the bone marrow. In a person with Leukemia, abnormal white cells (leukemic blast cells) are produced which cannot further produce normal white blood cells. In contrast to normal blood cells, Leukemia cells do not perish when they become old or damaged [10]. These cells can further produce more leukemic blast cells, which can out populate normal blood cells leading to issues like the risk of infections, difficulty in controlling bleeding.

Acute Leukemia is classified into two categories in accordance with French-American-British (FAB) classification: Acute Lymphoblastic Leukemia (ALL) and Acute Myelogenous Leukemia (AML). ALL affects lymphoid cells and develops rapidly [1]. There are nearly 25,000 children diagnosed with cancer in India every year and around 9000 of these have leukemia [11]. The early prediction of leukemia is of utmost importance in rendering the appropriate treatment. For classification, the two main analyses: a qualitative study of cell morphology and quantitative approach of counting the WBC. The qualitative approach is subject to the operator's abilities. In quantitative approach, the blood sample gets destroyed during analysis [2]. The whole process is time-consuming and expensive.

In this paper, we propose an automated system for classification of white blood cells from microscopic images. The system adopts methodologies that process the input blood image sample to extract features from the required components i.e. Nucleus of the WBC. The system is aimed towards the efficient and accurate classification of white blood cells as blast cells or normal cells.

PROPOSED SYSTEM

In this paper, we focus on the problem of Leukemia prediction by proposing a system which has the following sub-systems: Image Preprocessing, Feature Extraction and Cell Classification [7]. The system performs image processing and therefore just requires an image of a blood sample but not a blood sample and hence is suitable for low cost, standard-accurate and remote diagnostic systems. The input blood sample is fed into the system.

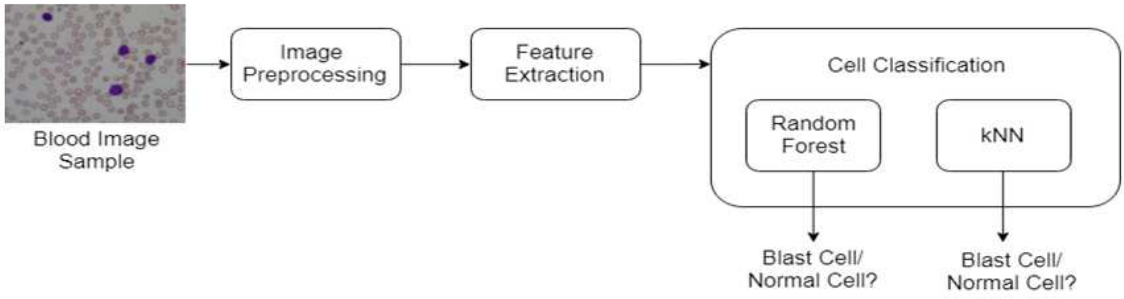

Figure 1: Block Diagram of the Proposed System

Nature of white blood cells provides worthwhile information in the process of medical diagnosis. White blood cells are divided into five categories: Neutrophil, Eosinophil, Basophil, Monocyte and Lymphocyte. Since the growth of lymphocyte being responsible for leukemia, lymphocytes are identified in the pre-processing module of the system [2]. The geometric and statistical features of these cells are extracted [7]. Prediction of leukemia is achieved by feeding the extracted features to the cell classifier to classify it as blast cell or normal cell.

Image Acquisition

Input blood images are obtained from a public and free dataset of microscopic images of blood samples (ALL-IDB) maintained by Dr. Fabio Scotti, specifically designed for the evaluation and the comparison of algorithms for segmentation and image classification [8]. The images of the dataset have been captured with an optical laboratory microscope coupled with a Canon Power Shot G5 camera [12]. Two types of datasets are available: ALL-IDB1, which can be used both for testing segmentation capability of algorithms, as well as the classification systems and image pre-processing methods, and ALL-IDB2, which has been designed for testing the performances of classification systems [8].

Image Preprocessing

The image pre-processing module focuses on the enhancement of the blood sample image and selection of the white blood cells by removing the other blood components viz. Red Blood Cells (RBC) and platelets. The algorithm for pre-processing and segmentation of lymphocytes is as follows [4]:

- **Input:** RGB color blood sample image
- **Output:** Binary images of individual lymphocytes
- Convert the color image to grayscale image.
- Perform histogram equalization of grayscale image to enhance contrast (A).
- Apply linear contrast stretching to adjust image intensity (B).

- Obtain image I1 = B + A to brighten all other image components except cell nucleus.
- Obtain image I2 = I1 – A to highlight the entire image objects along with cell nucleus.
- Obtain image I3 = I1 + I2 to remove all other components of blood.
- Apply 3x3 median filter to reduce the noise and preserve edges.
- Apply k-Means clustering to convert grayscale image to binary image.
- Perform morphological opening to remove small pixel groups
- Apply SAGAP algorithm to extract the connected components from the binary image [5].
- This method results in the segmentation of lymphocytes with good accuracy.

Feature Extraction

Feature extraction refers to the transfer of the input data i.e. Blood sample images, into a different set of features. In the proposed system, the feature extraction module incorporates algorithms which are used to detect and isolate various desired portions or shapes (features) of an image. This module is most critical because the accuracy of this process determines the performance of the classification algorithm used. The features extracted are geometric features viz. area, perimeter and compactness since the shape of the nucleus is an important feature for differentiation of blast cells [1]. These features are extracted from the binary image of lymphocyte. The Statistical features viz. Standard deviation, variance, energy and entropy are extracted from the grayscale equivalent image of the cell [3].

Area: The area is estimated by counting the number of non-zero pixels in the binary image of lymphocyte.

Perimeter: The perimeter is measured by calculating the distance between successive boundary pixels [6].

Compactness: It is a dimensionless parameter, which is a function of area and perimeter

$$Compactness = \frac{4 * \pi * Area}{Perimeter^2}$$

Standard Deviation: describes how much the pixel value differs from the mean pixel value.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - U)^2}$$

Variance: measures how far each pixel in the set is from the mean

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - U)^2$$

Energy: describes how grey levels are distributed.

$$Energy = \sum_{g=1}^{N} P(g)^2$$

Entropy: a statistical measure of random values used to characterize the texture of the input image

$$Entropy = -\sum_{g=1}^{N} P(g)log_2(P(g))$$

Cell Classification

The features extracted in the Feature Extraction module are used by the Random Forest and The k-Nearest Neighbor (kNN) classifiers to classify the lymphocyte as blast cell or normal cell. Random Forest classifier is an eager learner that operate by constructing a multitude of decision trees at training time and outputting the class i.e. the mode of the classes [9]. It runs efficiently on large databases and is stable, particularly on high dimensional spaces. The k-Nearest Neighbor is a non-parametric classification method [13]. It is a lazy learner i.e. there is no learning phase, instead, the training data is "memorized" and is simple and scalable.

Random forest algorithm

- **Input:** Feature matrix (training data), test record
- **Output:** Class label of the test record
- Grow a forest of many trees.
- Grow each tree on an independent bootstrap sample (Sample N cases at random with replacement) from the training data.
- For each tree:
- Select 'm' variables at random out of all M possible variables (independently for each tree).
- Find the best split on the selected m variables.
- Grow the trees to maximum depth (classification).
- Vote the trees to get predictions for new data.

The performance of both the classification algorithms is analyzed to find the most suitable for the purpose of leukemia prediction.

EXPERIMENTAL RESULTS

The proposed technique for preprocessing the blood samples is applied to the images from the public dataset. The results of the preprocessing algorithm mentioned in section 4 applied to figure 2, are shown in figure 3, 4 and 5.

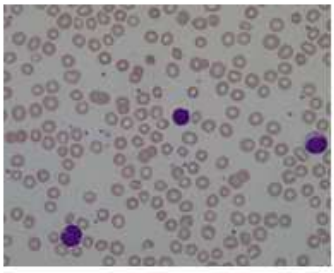

Figure 2: Original Image

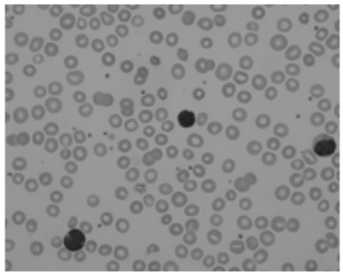

Figure 3: Image Converted to Grayscale

Figure 4: Image After Median Filtering

Figure 5: Individual Lymphocytes Extracted from the Blood Sample

The above images show the result of preprocessing module. Figure 4 shows the output after performing filtering to obtain only the lymphocytes and remove the other unwanted components. Figure 5 shows the individual lymphocytes that are extracted by applying SAGAP algorithm on segmented image. These features of these components are extracted and fed to the Random Forest and kNN classifiers.

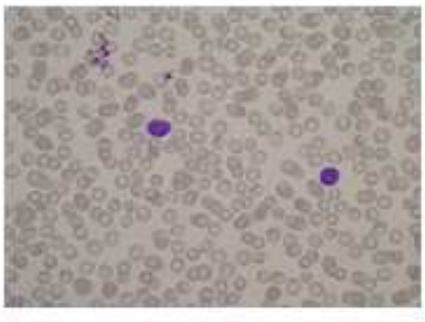

Figure 6: Original Blood Image Sample

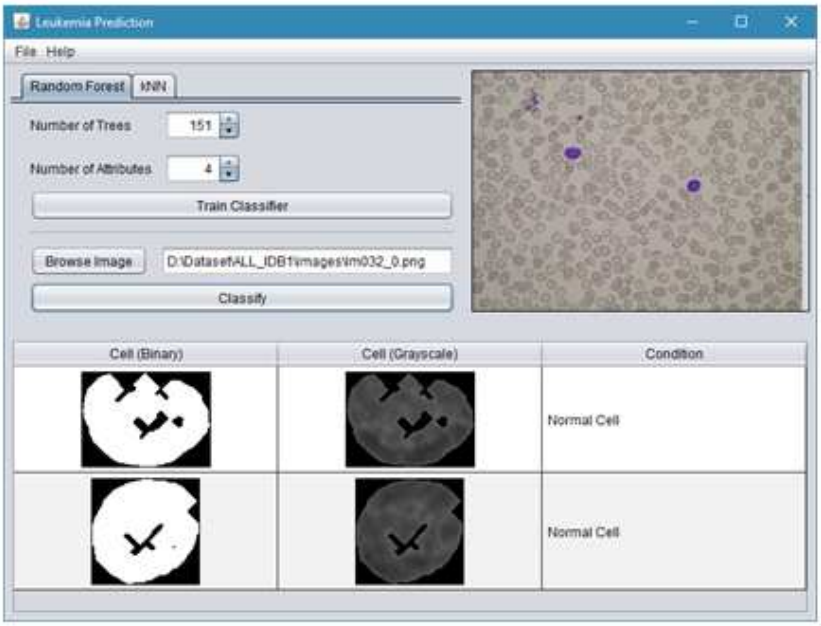

Figure 7: Output of Random Forest Classifier

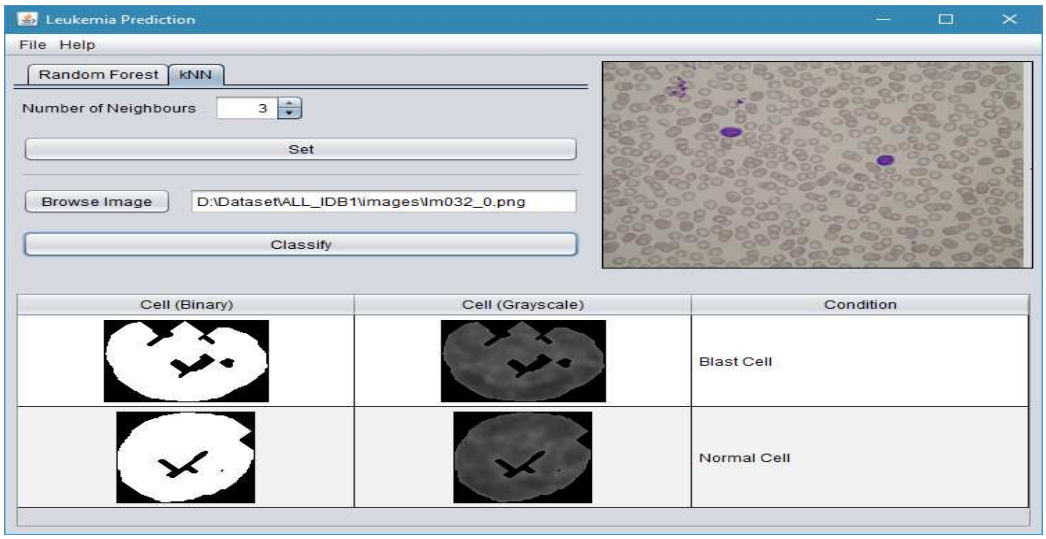

Figure 8: Output of KNN Classifier

Figure 7 and Figure 8 show the output of the proposed system on the blood sample shown in Figure 6. The blood sample shown in Figure 6 is from a healthy individual i.e. both the cells are normal cells. It can be seen from the output that Random Forest classifier accurately classifies both the cells as normal cells whereas the kNN classifier classifies one cell incorrectly as a blast cell. A detailed analysis of the performance of the classifiers is carried out based on the following parameters [8]:

- True positives (TP) – the number of elements correctly classified as positive by the test;
- True negatives (TN) – the number of elements correctly classified as negative by the test;
- False positives (FP) – type I error, the number of elements classified as positive by the test, but they are not;
- False negatives (FN) – type II error, the number of elements classified as negative by the test, but they are not;

- Sensitivity – the probability of correctly classifying elements with ALL. $Sensitivity = \frac{TP}{(TP+FN)}$
- Specificity – the probability of correctly classifying elements without ALL. $Specificity = \frac{TN}{(TN+FP)}$
- Classification error – total error in analysis. $Classification\ Error = FP + FN$

Table 1: Analysis of Random Forest and k Nearest Neighbor Classifiers

Parameter	Random Forest	k Nearest Neighbor
Accuracy	95 %	75 %
TP %	45 %	40 %
TN %	50 %	35 %
FP %	0 %	15 %
FN %	5 %	10 %
Misclassification %	5 %	25 %
Sensitivity %	90 %	80 %
Specificity %	100 %	70 %

Random forest algorithm was used to generate 151 trees with subsets of 4 attributes of the total 7 attributes. This configuration produced a remarkable accuracy of 95%. kNN algorithm produced the best accuracy i.e. 75% when k = 3.

CONCLUSIONS

This paper aims to achieve an automated system for prediction of leukemia in blood samples. The system incorporated various preprocessing techniques to ensure that noise and other unwanted elements are removed with a high degree of accuracy. The features that were most crucial in the prediction of Leukemia namely area, perimeter, compactness, standard deviation, variance, energy and entropy were selected to be used with the Random Forest and kNN classification algorithm. Random Forest, being an ensemble method produces better classification than a single classifier. It was found out that Random Forest is preferable than kNN for the requirements of our system as it provides an accuracy of 95%. Thus, the automated system developed using the mentioned methods would be helpful in early detection of the disease and thereby more opportunities for treatment.

The results of the study presented in this paper could motivate further studies increasing the accuracy of classification by considering additional features. Furthermore, techniques can be devised to improvise the segmentation of blood images where the cells are in contact with each other.

REFERENCES

1. *MinalD. Joshi, A.H.Karode, S.R.Suralkar, "White Blood Cell Segmentation and Classification to Detect Acute Leukemia", IJETTCS, e ISSN: 2278-6856, Volume 2, Issue 3, May - June 2013*
2. *V. Piuri, F. Scotti, "Morphological classification of blood leucocytes by microscope images", in Proc. of the 2004 IEEE Int. Conf. on Computational Intelligence for Measurement Systems and Applications (CIMSA 2004), Boston, MA, USA, pp. 103-108, July 12-14, 2004. ISBN: 0-7803-8341-9*
3. *Fauzia Kasmin, Anton Prabuwono, AziziAbdullah, "Detection of Leukemia in Human Blood Sample based on Microscopic Images: A Study", Journal of Theoretical and Applied Information Technology, e ISSN: 1817-3195Vol. 46 No.2, Dec 2012*

4. *H.T. Madhloom, S.A. Kareem, H. Ariffin, A.A. Zaidan, H.O. Alanzi, B.B. Zaidar, "An Automated White Blood Cell Nucleus Localization and Segmentation using Image Arithmetic and Automatic Threshold", Journal of Applied Sciences, e ISSN: 1812-5654, 2010*

5. *Prakash Kumar, SaumyaRanjanGiri, Ganesh Rama Hegde, KanchanVerma, "A Novel Algorithm to Extract Connected Components in a Binary Image of Vehicle License Plates", IJECCT 2012, Vol. 2*

6. *Eng. Ayman M Bahaa Eldeen Sadeq, Prof. Dr. Abdel-Moneim A. Wahdan, Prof. Dr.Hani M. K. Mahdi Faculty of Engineering, Ain Shams University, "Edge Detection Of Binary Images Using The Method Of Masks.", Vol. 35, No. 3, Sept 30, 2000.*

7. *Pratik Gumble, Dr. S.V. Rode, "Study and Analysis of Acute Lymphoblastic Leukemia Blood Cells Using Image Processing", IJIRCCE, e ISSN: 2320-9801, Vol. 5, Issue 1, January 2017*

8. *R. Donida Labati, V. Piuri, F. Scotti, "ALL-IDB: the acute lymphoblastic leukemia image database for image processing", in Proc. of the 2011 IEEE Int. Conf. on Image Processing (ICIP 2011), Brussels, Belgium, pp. 2045-2048, September 11-14, 2011. ISBN: 978-1-4577-1302-6.*

9. *Leo Breiman & Adele Cutler, Random Forests. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm*

10. *U.S. Department Of Health And Human Services, National Institutes of Health, What You Need To Know About™ Leukemia. https://www.cancer.gov/publications/patient-education/wyntk-leukemia/AllPages*

11. *International Agency for Research on Cancer, Population fact sheet: India. http://globocan.iarc.fr/*

12. *Fabio Scotti et al, Acute Lymphoblastic Leukemia Image Database for Image Processing. https://homes.di.unimi.it/scotti/all/*